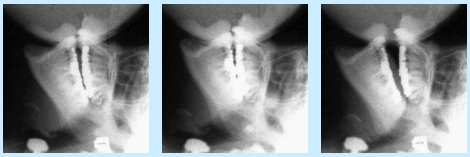
QUASI-AUTOMATIC EXTRACTION OF TONGUE MOVEMENT FROM A LARGE EXISTING SPEECH CINERADIOGRAPHIC DATABASE

Julie Fontecave and Frédéric Berthommier
Institut de la Communication Parlée, INP Grenoble, France
E-mail : fonte, bertho@icp.inpg.fr

Introduction

Extraction of geometrical information starting from a video observation of the motion in a speech context
Development of algorithms appropriate with the capture of biological movements without the use of markers and without a contour extraction technique
Analysis of non visible vocal tract movements (possible exploitation of a large amount of data)

Problem : very poor contrast and occlusion > Difficult manual processing image by image for extracting tongue shape (position) from static images. Motion must be taken into account to extract the geometrical features
Advantage : large database > high time redundancy of the movements (vocal tract gestures are pseudo-periodic)



Our semi-automatic extraction method of articulatory information =
a limited manual step + an automatic extraction of the geometrical information in the full database

1. Retro Marking :
Development of a semi-automatic marking method for video data

The manual processing of a few key images, associated with the automatic indexing of the full database according to the keys, provides a geometrical marking of the original images.

Video Database : Cineradiographic data of the Vocal Tract
5673 images (490*480 pixels) recorded at 25 im/sec, from 64 video sequences
(sentences pronounced by a female French Speaker)

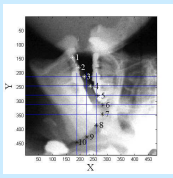
Video

575 (24*24 – 1) low frequency **DCT components** of each image, calculated on resized, centered and framed images (104*99 pixels)

100 Key Images chosen randomly among the full database

DCT

Manual marking of the key images
Description of the position of the tongue contour with only 10 points (12 degrees of freedom, d.o.f.)



INDEX
calculated for each image by assigning the index of the nearest key image
similarity measure = Euclidian distance on the video features

Geometrical features of the whole database
Geometrical marking of the original images by association via the indexing of the geometrical information available on the key images only

Geometry

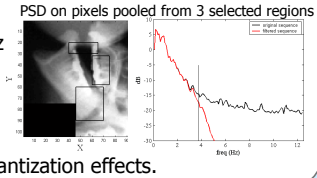
2. Reconstruction of the geometrical information across time

Reduction of the baseline reconstruction error by restoring continuity
Observation of the effect of error reduction operators on intermediate representations : Principal Component Analysis (PCA) on video features (DCT components) and geometrical features (marked points coordinates) of key images

Temporal filtering of geometrical features

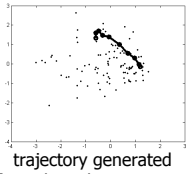
The video components frequency is about 3.75Hz as for the geometrical information. A temporal filtering is applied on the sequence of geometrical features to reduce the quantization effects.

PSD on pixels pooled from 3 selected regions



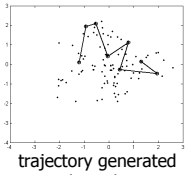
Neighborhood averaring

PCA Video Space | PCA Geometrical Space



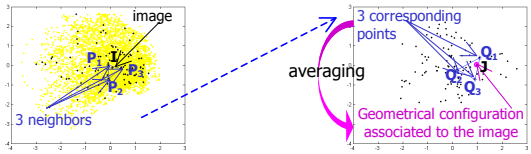
trajectory generated from the video sequence

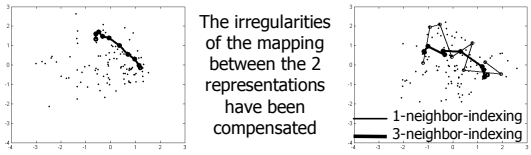
Observation of **trajectories** : Irregularity of the relationship between video and geometry



trajectory generated via the indexing

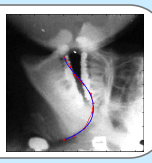
Attenuation of the geometrical trajectory discontinuities by **averaging the geometrical configurations of the 3 neighbors** taken in the video space





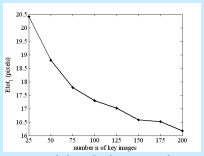
The irregularities of the mapping between the 2 representations have been compensated

A quite irregular (red) tongue contour is obtained by connecting the 10 points. A **Spline Interpolation** by a 5-degree polynomial on the points improves the (blue) geometrical configuration for each frame.



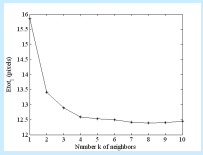
3. Evaluation

Starting from an expert manual marks on 200 images (100 key images and 100 test images), evaluation of the **reconstruction error RMS** (root mean square) in pixel for 490*480 images among the pair of geometrical features, the key one and the test one : evolution of the RMS error mean value on the 12 d.o.f. according to the different parameters



model including simple indexing and temporal filtering

Why 100 key images ?
Compromise between the reconstruction error rate and the cost of the manual processing.

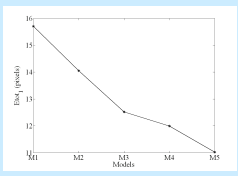


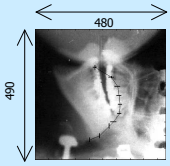
model without temporal filtering

Number *k* of neighbors
Increasing *k* from 1 to 4 provides a great error reduction (but no supplementary gain for *k*>4)

Models are built using combinations of the 3 error reduction methods. These are **complementary** : the reconstruction error is reduced gradually.

Methods \ Models	M1	M2	M3	M4	M5
Neighborhood size	1	1	4	4	4
Temporal filtering	-	+	-	+	+
Spline Interpolation	-	-	-	-	+





Error seen d.o.f. by d.o.f. for M4 model on the marks of one key image
It is uniformly distributed along the tongue contour.

Conclusion

- A limited manual processing step
- An automatic "retro-marking" treatment (taking a few minutes)
- The tongue movement is retrieved (with an error of a few pixels)

Prospects

- Extensions to lips, soft palate… > Vocal tract configurations and speech synthesis
- Correlation between lips and tongue, inversion

Acknowledgments : We thank P. Perrier and R. Sock for providing the video (recorded in 1985 by F. Wioland) digitized in the context of the CNRS project of valorization of cineradiographic data